Toward efficient Bayesian solution of inverse problems

Youssef Marzouk
Sandia National Laboratories
Livermore, CA 94551 USA

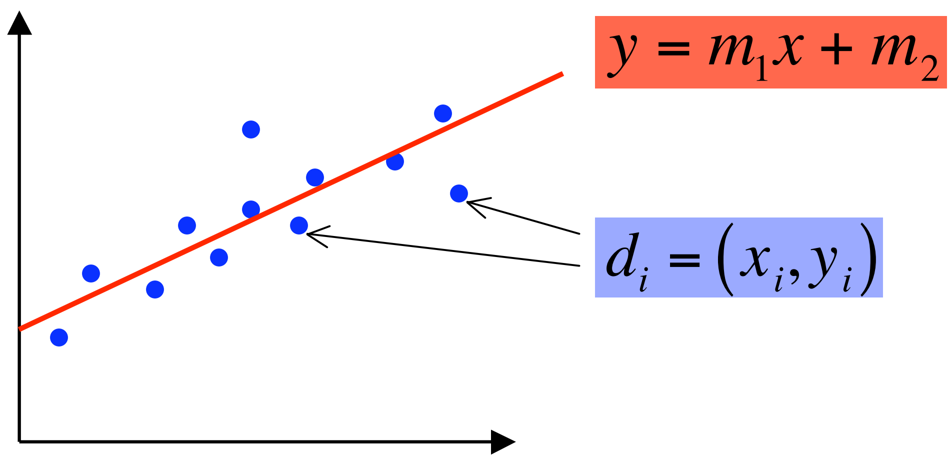
jointly with Habib Najm (SNL), Larry Rahn (SNL)

support from: SNL LDRD, DOE Basic Energy Sciences

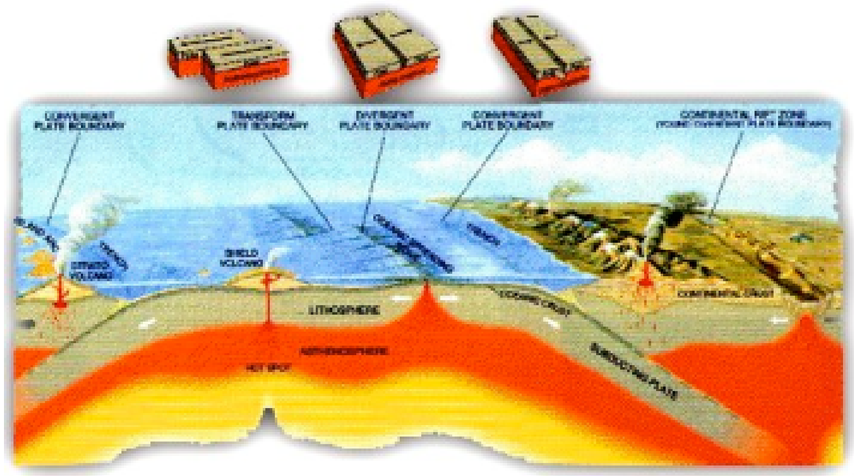
Recognizing inverse problems…

- How to relate (indirect) *observations* to physical *parameters* and *models?*

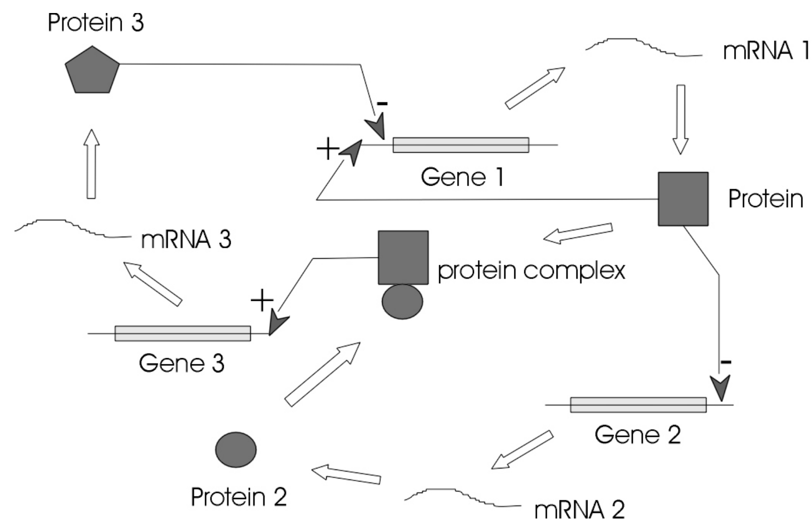
simple example =
linear regression
(solve with least squares)



Inverse problem examples



geophysics (seismic profiling, prospecting)



[J. Gebbert, U Köln]

building models of gene regulatory networks

medical imaging & tomography

+ source inversion
(security, environment)

Inverse problems

$$G_\theta(m) = d$$

"forward" operator

observed data

model parameters

Given a set of data d, estimate m
(or, estimate m and G_θ)

∴ Formalize the process of inference. "Finding unknown *causes* based on their *effects* [Alifanov]"

Inverse problems

- Why are they difficult?
  - G^{-1} often non-local, non-causal.

$\Rightarrow$ Classically ill-posed:

1 No solution may match the data *(existence)*

- linear case, $G \in R^{m x n}$: $\exists$ a non-trivial *data nullspace* $N(G^T)$

2 Many solutions may match the data *(uniqueness)*

- linear case: $\exists$ a non-trivial *model nullspace* $N(G)$; more likely when data is **sparse** or **degenerate** relative to *dim(m)*

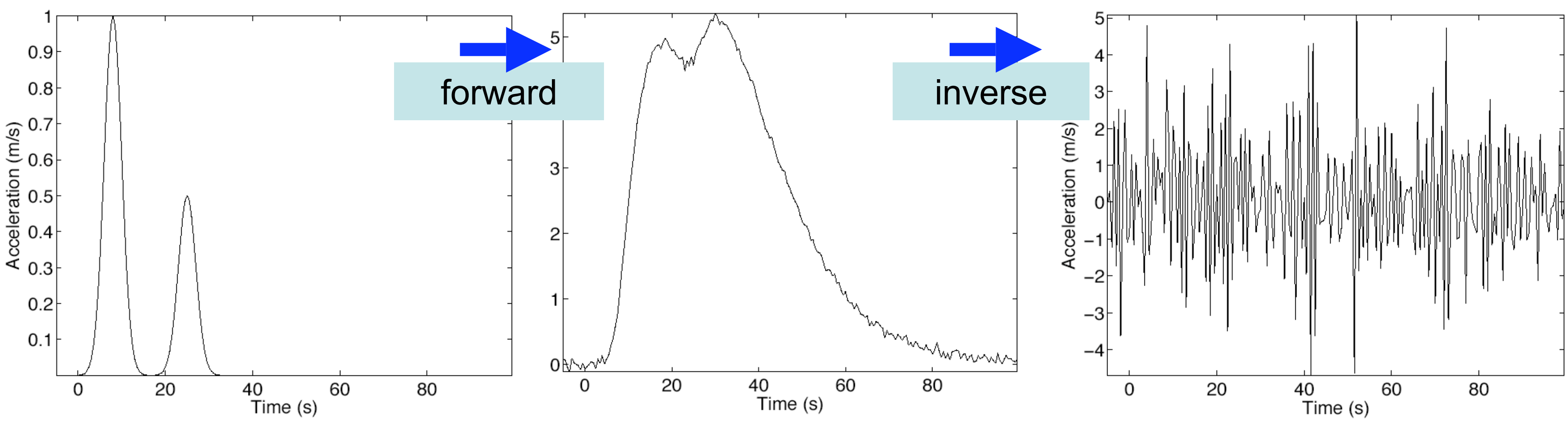
3 Ill-conditioning or *instability:* Small changes in data d can lead to large changes in m

- linear case: singular values $\sigma_i(G)$ decay rapidly towards zero

$\Rightarrow$ result: sensitivity to noise

Noise and ill-conditioning

- Example: de-convolve ground acceleration from seismometer output [from ABT04]



true model
(acceleration profile)

seismometer data +
uncorrelated 5% rms
noise

inverse solution
(not a very good solution)

Deterministic approaches

- Usual approach: regularization + optimization
- Regularization: impose smoothness, positivity, maximum entropy, etc…
- Example: Tikhonov-type regularization

$$\text{minimize} \quad J = \left\|G(\mathbf{m}) - \mathbf{d}\right\|_2^2 + \boxed{\alpha \left\|\mathbf{Lm}\right\|_2^2}$$

e.g., a roughening matrix L

- Drawbacks:
  - How to choose **L**, α, etc? Regularization can be somewhat arbitrary.
  - Regularization introduces *bias*, destroys consistency.
  - No meaningful uncertainty/confidence intervals on the resulting *m.*

Outline

1 Inverse problems

2 Bayesian solution of inverse problems

- Formulation; Bayesian inference
- Results: source inversion under transient diffusion

3 New computational tools for Bayesian inversion

- Spectral representations of stochastic processes
- Polynomial chaos in Bayesian inference
- Results: accelerated MC and MCMC simulation

4 Extensions

Bayesian inference for IPs

- Let the model *m* be a random variable
- Bayes' theorem:

likelihood function L(m)

prior distribution

$$\underbrace{p(m|d)}_{\text{posterior distribution}} = \frac{\overbrace{p(d|m)}^{\text{likelihood function } L(m)}\,\overbrace{\rho(m)}^{\text{prior distribution}}}{\underbrace{\int p(d|m)\rho(m)\,dm}_{\text{evidence (here just a normalizing const)}}}$$

posterior distribution

evidence
(here just a normalizing const)

- Compared to classical approaches:
  - Not just a single value for *m*, but a probability density
    $\therefore$ **posterior = a COMPLETE description of uncertainty**
  - Additional information incorporated through the prior (expert judgment, additional experiments, physical constraints, etc…)
  - No regularization parameter *per se*

Likelihoods, priors, & hyperparameters

- Common shorthand for Bayes theorem:

$$\pi_{m|d}(m) \propto p(d|m)\, p_m(m)$$

- **Likelihood** function: $L(m) \equiv p(d|m)$
  (how well does the model support the data?)
  - *Example:* deterministic forward problem G(m); uncorrelated additive measurement + model errors $\eta_i \sim p_\eta(\xi)$

$$d_i = (G(m))_i + \eta_i \quad \rightarrow \quad L(m) = \prod_i p_\eta\left((G(m))_i - d_i\right)$$

  - Common choice: $p_\eta = N(0,\sigma^2)$
  - Alternate interpretation:

$$(d_{true} + \eta) \sim p_d(d) \quad \rightarrow \quad L(m) = p_d(G(m))$$

Likelihoods, priors, & hyperparameters

- **Prior** $p_m(m)$ comes from physical constraints, additional knowledge; can be ***uninformative***.
- Hyperparameters ϕ: what if we don't know some aspects of the noise/priors:
  - ex: $p_\eta = N(0,\sigma^2)$, σ^2 unknown

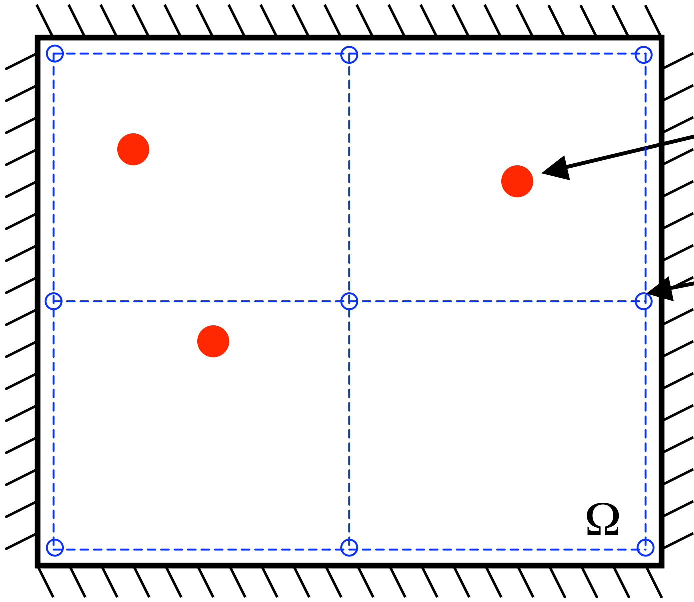
$$p(m,\phi|d) \propto p(d|m,\phi)p(m|\phi)p(\phi)$$

- The **posterior density** $\pi(m) \equiv p(m|d)$ IS the full Bayesian solution to the inverse problem!
- Computational challenge: how to extract information from the unnormalized posterior density?

$$E[f] = \int f(m)p(m|d)dm$$

- **What if the forward model is expensive?**

Source inversion— a model problem



N sources, each described by parameters $m = \{\chi_i, s_i, \sigma_i, \tau_i\}_{i=1\ldots N}$

Data from M sensors on a regular grid; $d = \{T_{t1}, T_{t2}, \ldots\}_{i=1\ldots M}$

$\Omega = [0,1]\times[0,1]$

$$\frac{\partial T}{\partial t} = \nabla_{\vec{x}}^2 T + \sum_i^N \frac{s_i}{2\pi\sigma_i^2} \exp\left(-\frac{\left|\vec{\chi}_i - \vec{x}\right|^2}{2\sigma_i^2}\right)\left[1 - H\left(t - \tau_i\right)\right]$$

$$\nabla T \cdot \hat{n} = 0 \; on \; \partial\Omega, \quad T(\vec{x},0) = 0$$

→ forward problem $G(m)=d$

Source inversion

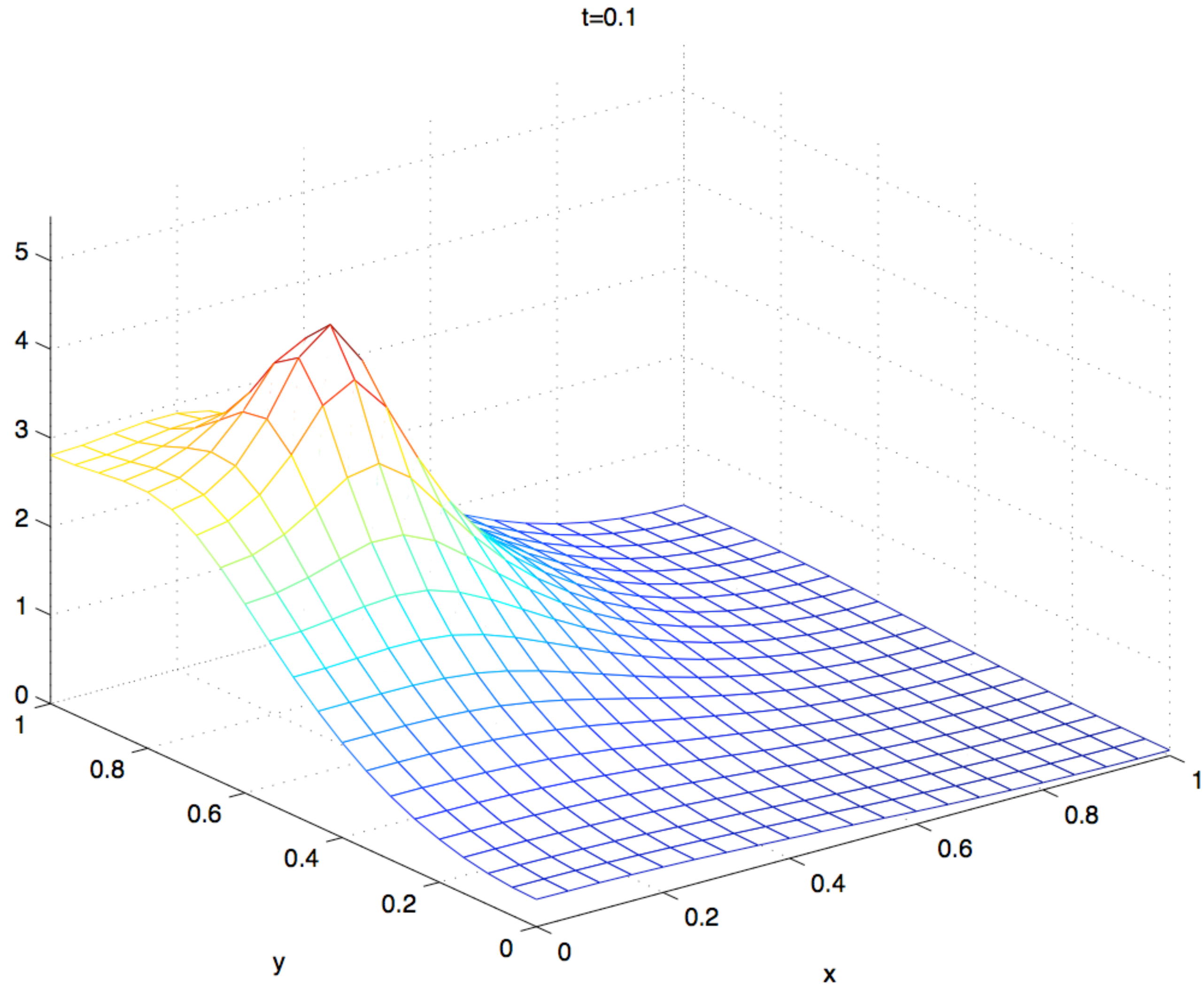
- To demonstrate, make some simplifications:
  - Consider only one source
  - Fix the source strength s_i, Gaussian width σ_i, and shutoff time τ_i
  - Goal: infer the source location $\chi = (x,y)$ from a small set of noisy measurements

  $\therefore$ This yields a 2-D posterior we can ***visualize***…

- Measurement **noise/error:** $\eta_i \sim N(0,0.2)$
- **Priors:** $(x,y) = (m_0,m_1) \sim U(0,1)$

Forward simulation

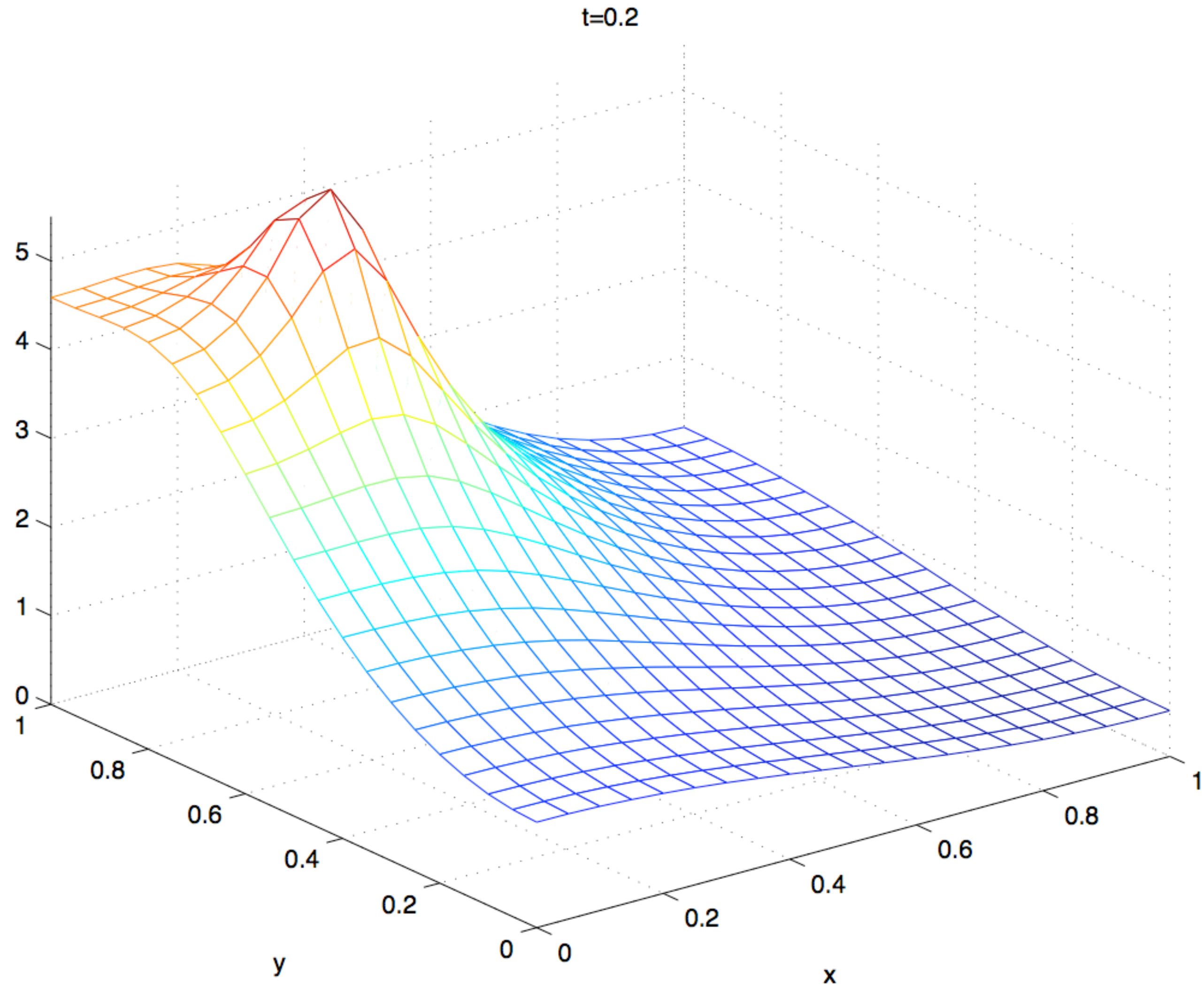
t=0.10



Source at *(x,y)* = *(0.25, 0.75)*; active for $t \in [0,0.2]$

Forward simulation

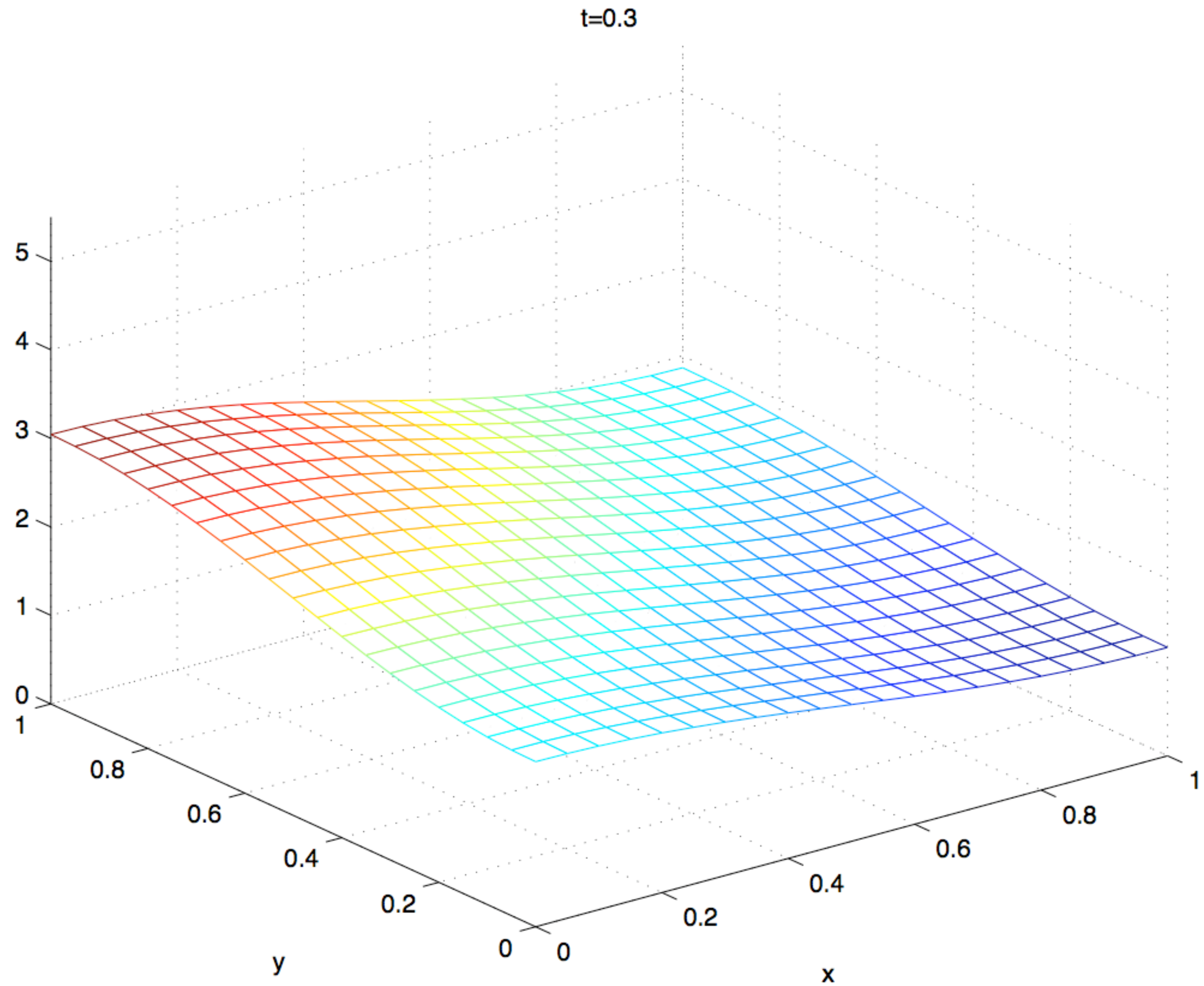
t=0.20



Source at $(x,y) = (0.25, 0.75)$; active for $t \in [0,0.2]$

Forward simulation

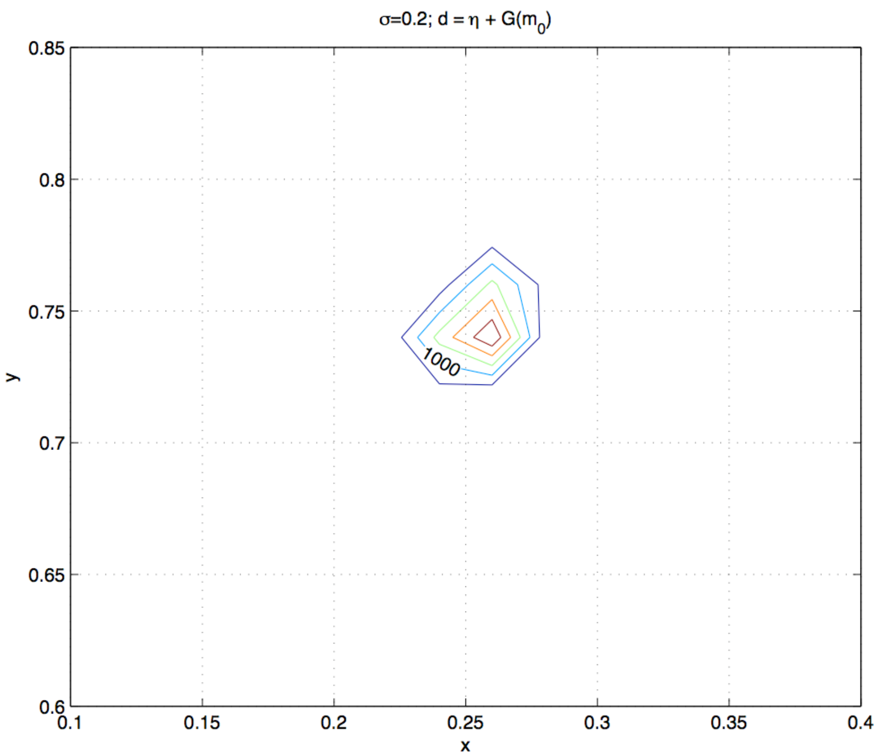
t=0.30



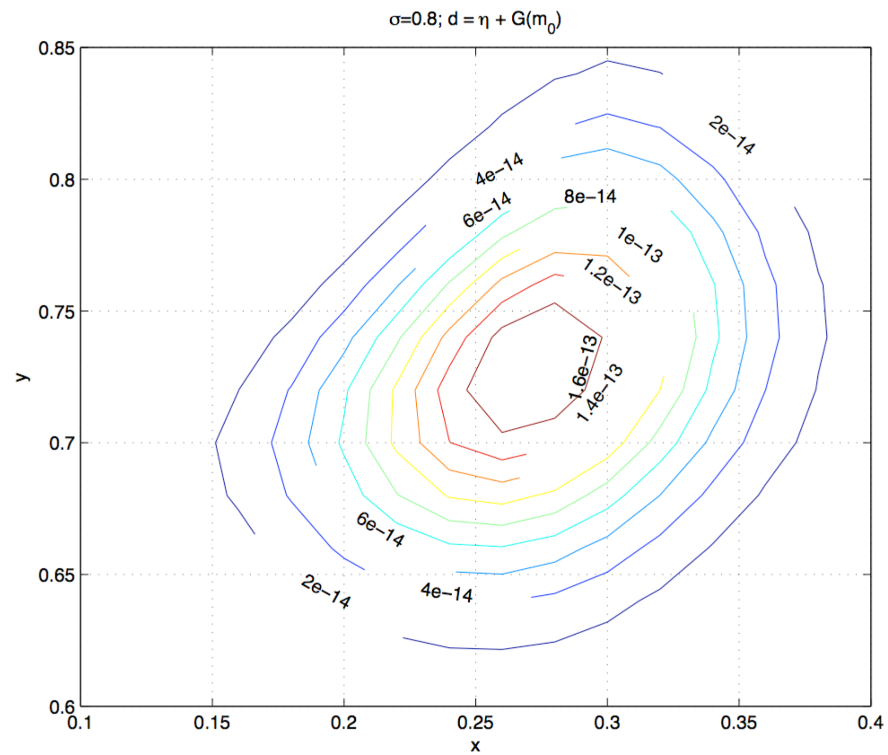
Source at $(x,y) = (0.25, 0.75)$; active for $t \in [0,0.2]$

Posterior density

- 3×3 grid of sensors; measure at $t = \{0.1, 0.2, 0.3\}$



noise $\eta \sim N(0, 0.2)$

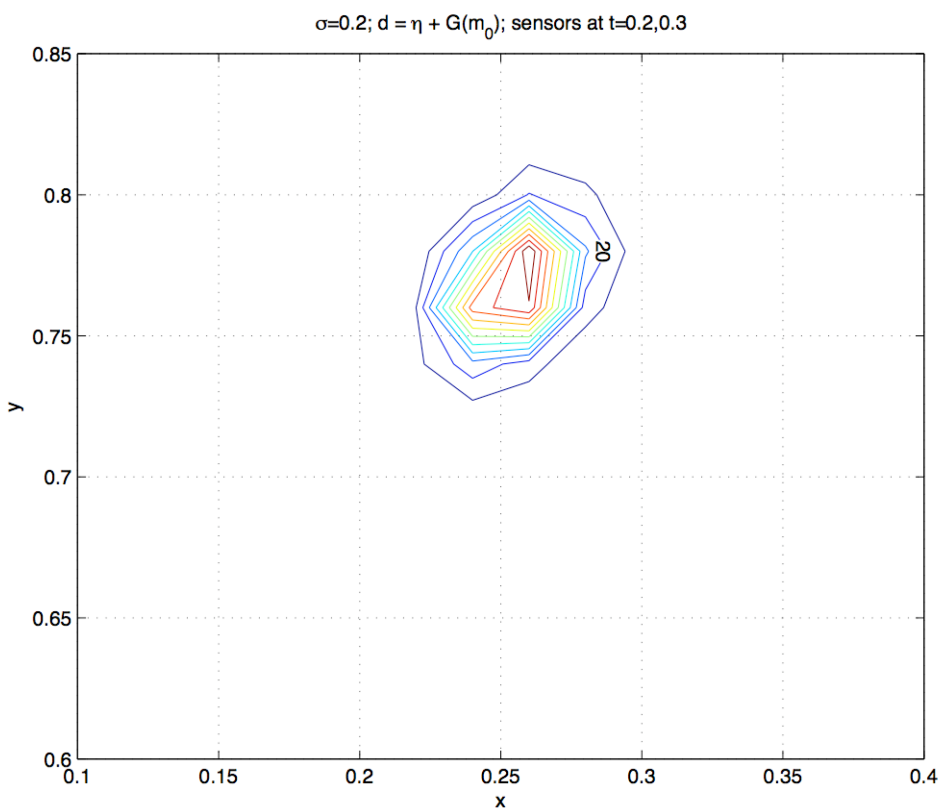


noise $\eta \sim N(0, 0.8)$

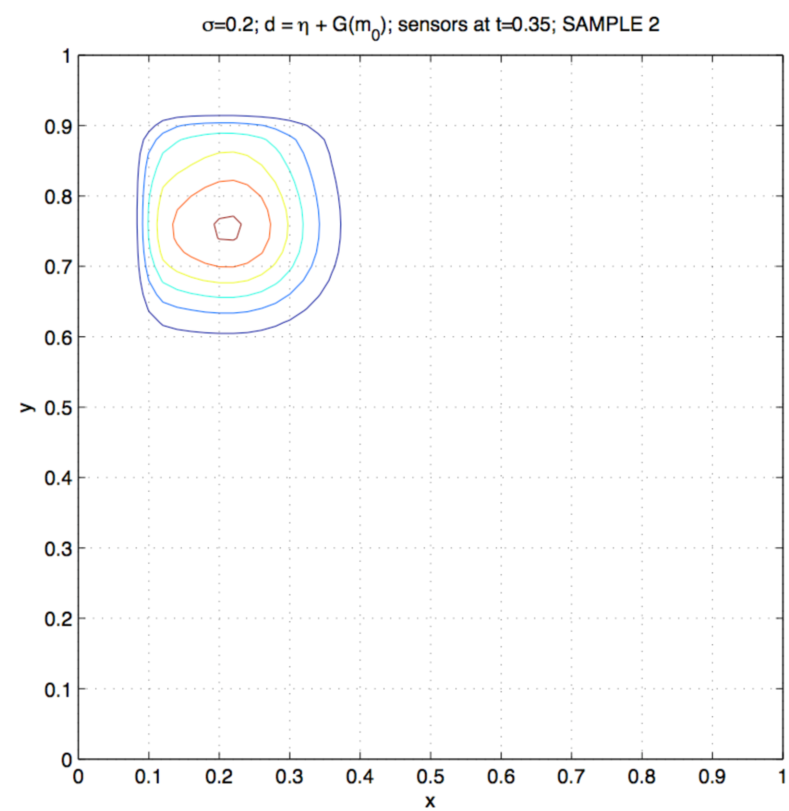
Posterior density

- Remove data, use more distant measurement times—make the problem more ill-conditioned.

$\Rightarrow$ broadens the posterior



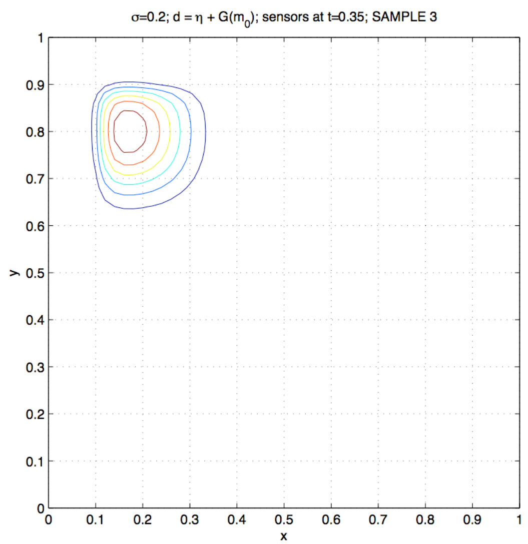
measure at $t = \{0.2, 0.3\}$;
$\eta \sim N(0, 0.2)$

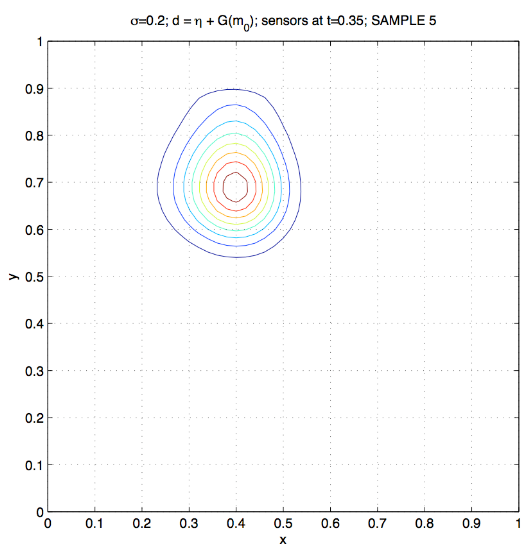


measure at $t = \{0.35\}$;
$\eta \sim N(0, 0.2)$

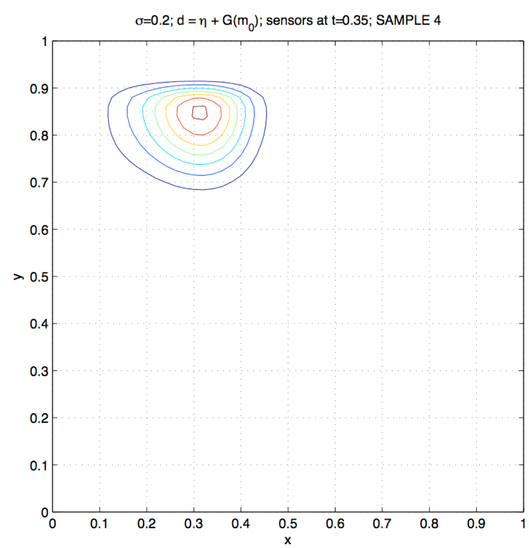
Posterior density

- Ill-conditioning ⇒ greater sensitivity to data (noise) realization



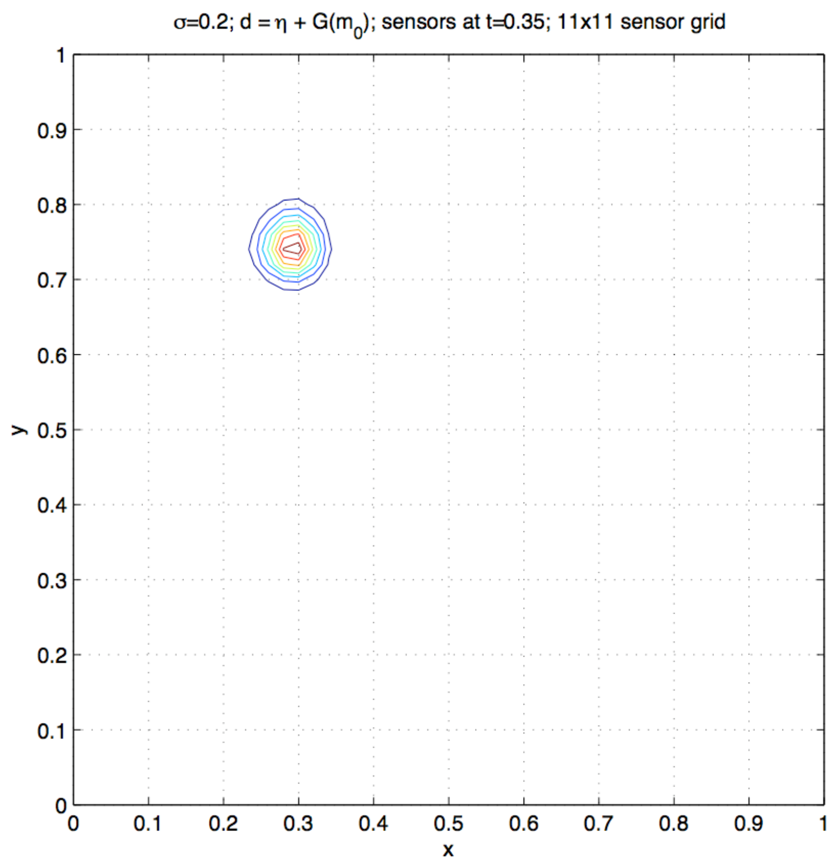


ALL: measure at $t = \{0.35\}$;
$\eta \sim N(0, 0.2)$



Posterior density

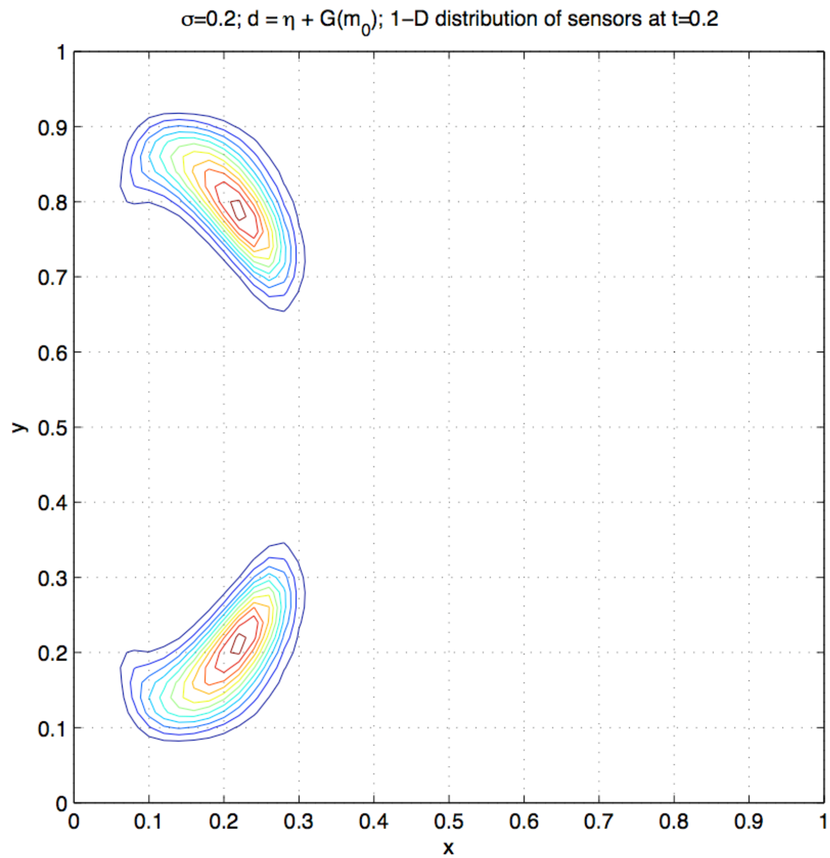
- Add more sensors → more precise knowledge; reduce ill-conditioning



measure at $t = \{0.35\}$;
$\eta \sim N(0, 0.2)$, 11×11 sensor grid

Posterior density

- **Non-unique solutions:**
  What if we had only a 1-D array of sensors? Place 3 sensors along the *y=0.5* line:



measure at $t = \{0.2\}$;

$\eta \sim N(0, 0.2)$, sensors at (x,y) = (0,0.5), (0.5,0.5), (1.0,0.5)

Outline

1 Inverse problems

2 Bayesian solution of inverse problems

- Formulation; Bayesian inference
- Results: source inversion under transient diffusion

3 Computational tools for Bayesian inversion

- Spectral representations of stochastic processes
- Polynomial chaos in Bayesian inference
- Results: accelerated MC and MCMC simulation

4 Extensions

Computational tools for Bayesian inference

- **Real (i.e., high-dimensional) problems—**what information to extract from the posterior?
  - Posterior means, variances, higher moments:

$$E_\pi[f] = \frac{I[f]}{I[1]} = \frac{\int f(m)\pi(m)\,dm}{\int \pi(m)\,dm}$$

  - Correlations, e.g., Cov(m_i, m_j)
  - Marginal distributions $p(m_i)$
  - Posterior “movie” (draw samples from the posterior)
- How to do this effectively?
  - Quadrature: $N_{evals} = O(n^d)$, prohibitive for large d.
  - Cubature (“sparse quadrature”): somewhat better scaling
  - **Sampling: Monte Carlo, Markov chain Monte Carlo (MCMC)**
- Challenge: posterior evaluations are expensive (forward problem)

Spectral rep'n of random variables

- Let (Ω, U, P) be a probability space, $X : \Omega \to R$ a square-integrable random variable. Then

$$X(\omega) = a_0\Gamma_0 + \sum_{i_1=1}^{\infty} a_1\Gamma_1\left(\xi_{i_1}\right) + \sum_{i_1=1}^{\infty}\sum_{i_2=1}^{i_1} a_{i_1 i_2}\Gamma_2\left(\xi_{i_1},\xi_{i_2}\right) + \cdots$$

$$= \sum_{k=0}^{\infty} \hat{a}_k \Psi_k\left(\xi_{i_1},\xi_{i_2},\ldots\right)$$

  - where $\left\{\xi_i(\omega)\right\}_{i=1}^{n}$ are orthonormal i.i.d. random variables
  - and $\left\{\Psi_k(\xi)\right\}$ are orthogonal multivariate polynomials:

$$\left\langle \Psi_i, \Psi_j \right\rangle = \int_{\Omega} \Psi_i(\xi)\Psi_j(\xi)\,dP(\omega) = \delta_{ij}\left\langle \Psi_i^2 \right\rangle$$

= a **polynomial chaos** expansion (PCe)

Spectral rep'n of random variables

- Many families of polynomials + distributions (Hermite + Gaussian, Legendre + Uniform, …)
- Truncate expansion at order *p* $\{\Gamma_0,\ldots,\Gamma_p\}$ and dimension *n* $\{\xi_1,\ldots,\xi_n\}$

$$\Rightarrow \{\Psi_k(\xi)\}_{k=1}^{P} \text{ where } P+1 = \frac{(n+p)!}{n!\,p!}$$

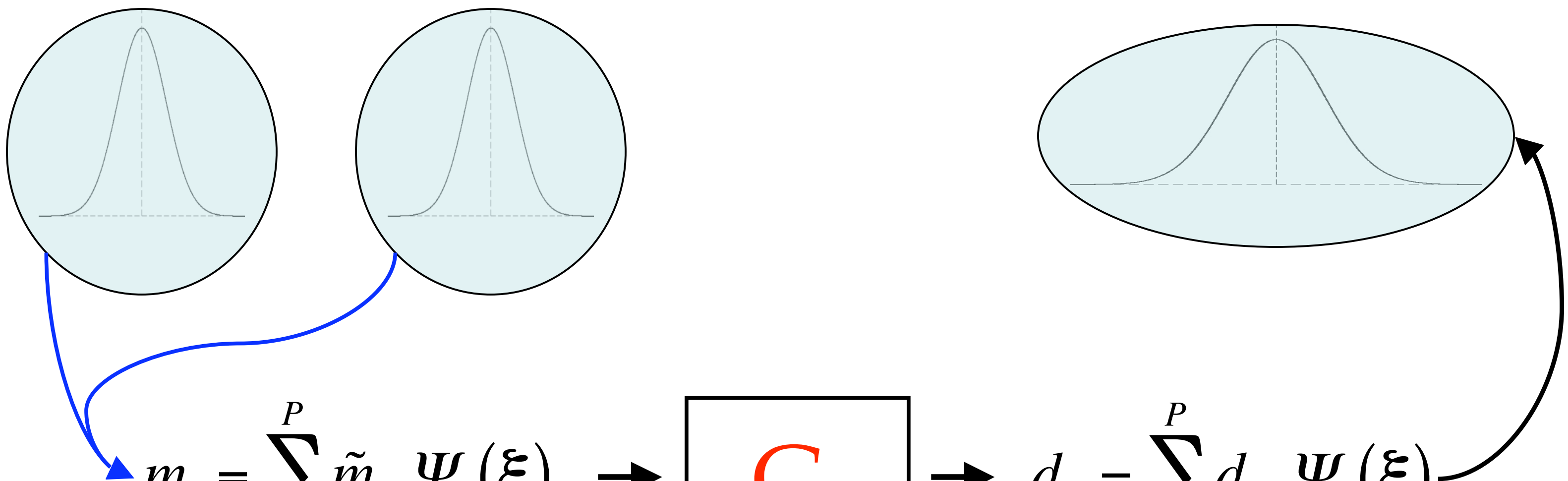
- Orthogonality: Galerkin projection determines spectral coefficients

$$g_k = \frac{\langle G(X)\Psi_k \rangle}{\langle \Psi_k^2 \rangle}$$

- Pseudo-spectral construction & other approaches for non-polynomial funcs; implemented in a library for "stochastic arithmetic."
- Primarily used in **uncertainty quantification:** structural, thermofluid, chemical systems [Ghanem, LeMaitre, Najm, Karniadakis]

PCe in Bayesian inference

- Write a PCe for ***m*** ~ **prior***:*



$$m_i = \sum_{k=0}^{P} \tilde{m}_{i,k} \Psi_k(\xi) \rightarrow \boxed{G} \rightarrow d_j = \sum_{k=0}^{P} d_{j,k} \Psi_k(\xi)$$

→ d_{jk}: spectral representation of the output of the forward model (compute **once!**)

PCe in Bayesian inference

- Draw samples $\xi^{(j)}$ from the distribution of ξ:
  - ***m(ξ)*** is thus sampled from its prior
  - Integrate over the posterior **without** repeated forward solutions:

$$I[f] = \int f(m)L(m)p_m(m)dm$$
$$\approx \frac{1}{N}\sum_{j=1}^{N}\left[f\left(m(\xi^{(j)})\right)\prod_i p_\eta\left(d_i - d_{i,PC}(\xi^{(j)})\right)\right]$$

- More generally, this corresponds to a change of variables ***m*** = ***g****(ξ)*:

$$\int_M f(m)L(m)p_m(m)dm \quad = \quad \int_{\tilde{\Xi}} f\left(g(\xi)\right)L\left(g(\xi)\right)p_m\left(g(\xi)\right)\left|\det(Dg)\right| d\xi$$

where ***g*** is a diffeomorphism mapping $\tilde{\Xi} \subseteq \Xi$ to the range of ***m***

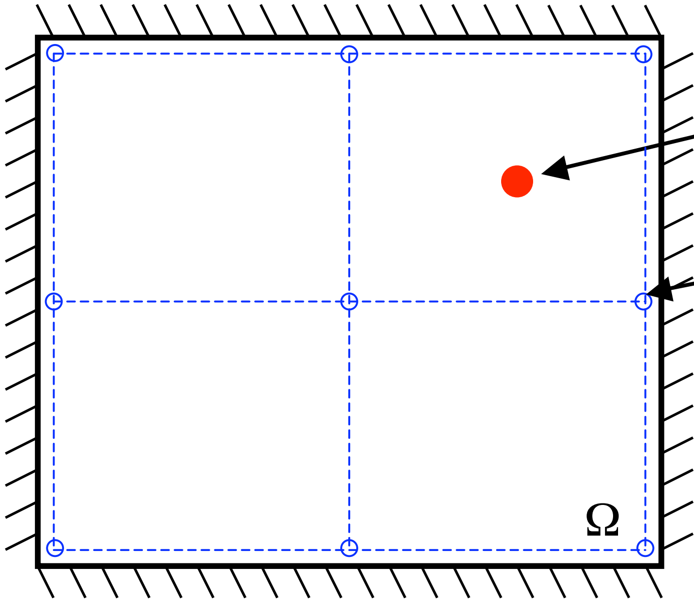
PCe in Bayesian inference

- Computational efficiency—partition the range of ***m*** into non-overlapping sets M^i***:***

$$p_m^i(m) = \begin{cases} p_m(m) & m \in M^i \\ 0 & m \notin M^i \end{cases}$$

Put $\boldsymbol{m}=\boldsymbol{g}^{\boldsymbol{i}}(\xi)$ on each subdomain.
Partitioning can be *adaptive* [LeMaître 2004; extends to wavelets…].

Source inversion



source described by parameters $m = \{\chi_i\}$, active for $t \in [0,0.2]$

Data from M sensors on a regular grid; $d = \{T_{t1}, T_{t2}, \ldots\}_{i=1\ldots M}$

$\Omega = [0,1]\times[0,1]$

$$\frac{\partial T}{\partial t} = \nabla_{\vec{x}}^2 T + \sum_i^N \frac{s_i}{2\pi\sigma_i^2}\exp\left(-\frac{\left|\vec{\chi}_i - \vec{x}\right|^2}{2\sigma_i^2}\right)\left[1 - H\left(t - \tau_i\right)\right]$$

$$\nabla T \cdot \hat{n} = 0 \ on \ \partial\Omega, \quad T(\vec{x},0) = 0$$

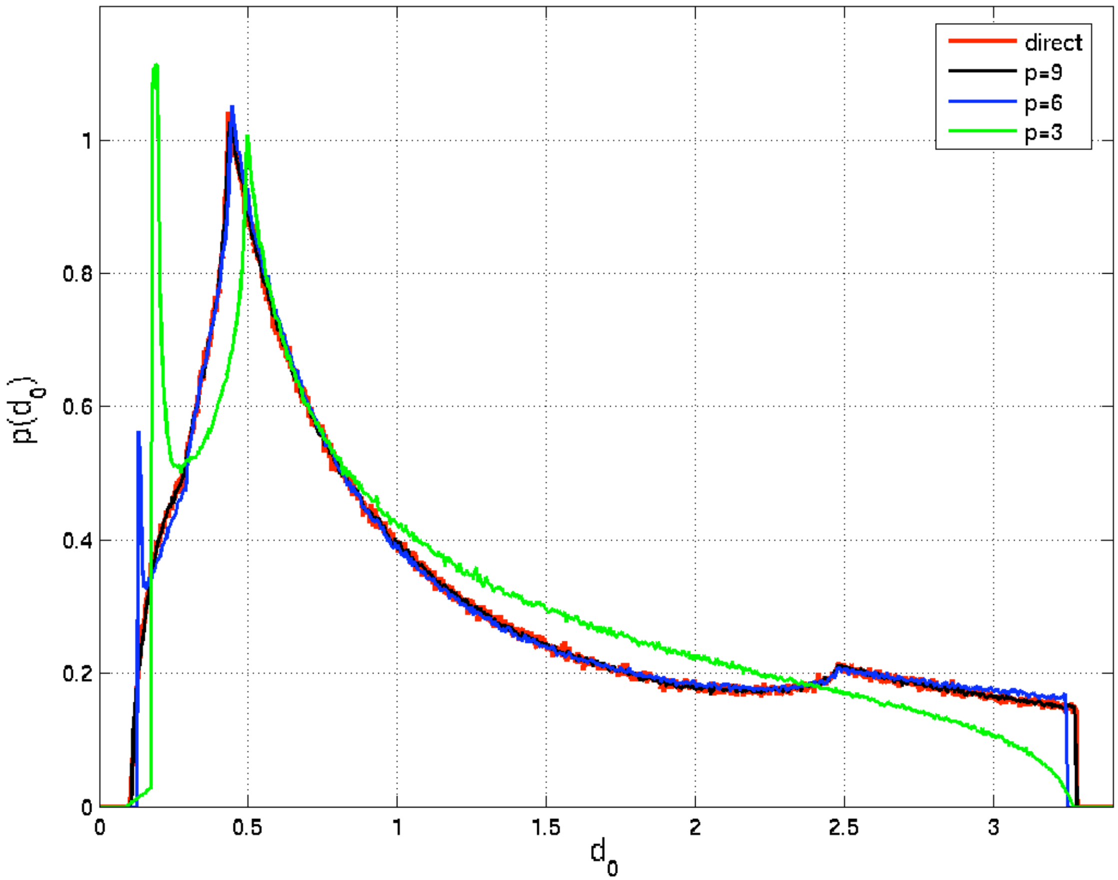
→ Measurement **noise/error:** $\eta_i \sim N(0,0.2)$

Priors: $(x,y) = (m_0,m_1) \sim U(0,1)$

⇒ Partition the support of the prior into 4 quadrants; solve the stochastic spectral forward problem on each domain.

Pdfs at measurement points

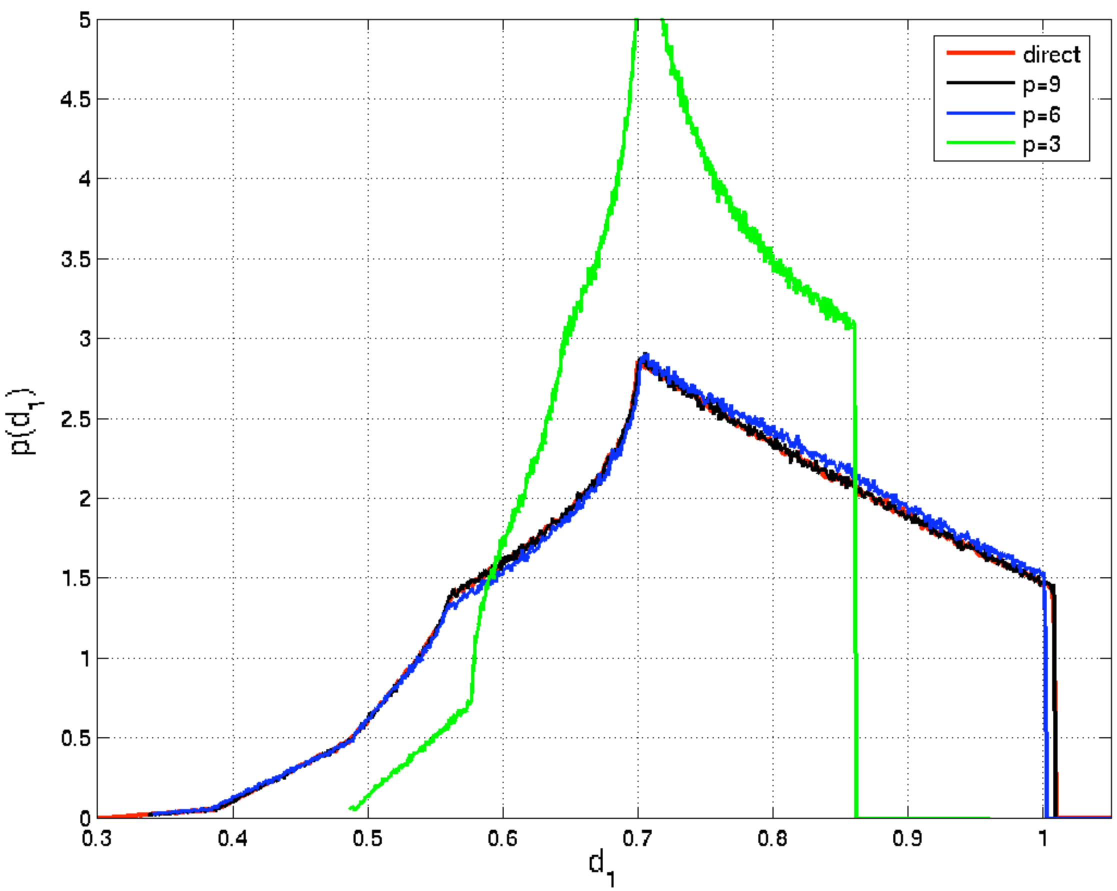
- Predicted value of the scalar field at *(x,y)* = *(0,0)*; ***t* = 0.05**
- Convergence with respect to **order** *p*



- Prior uniform on lower left quadrant of physical domain

Pdfs at measurement points

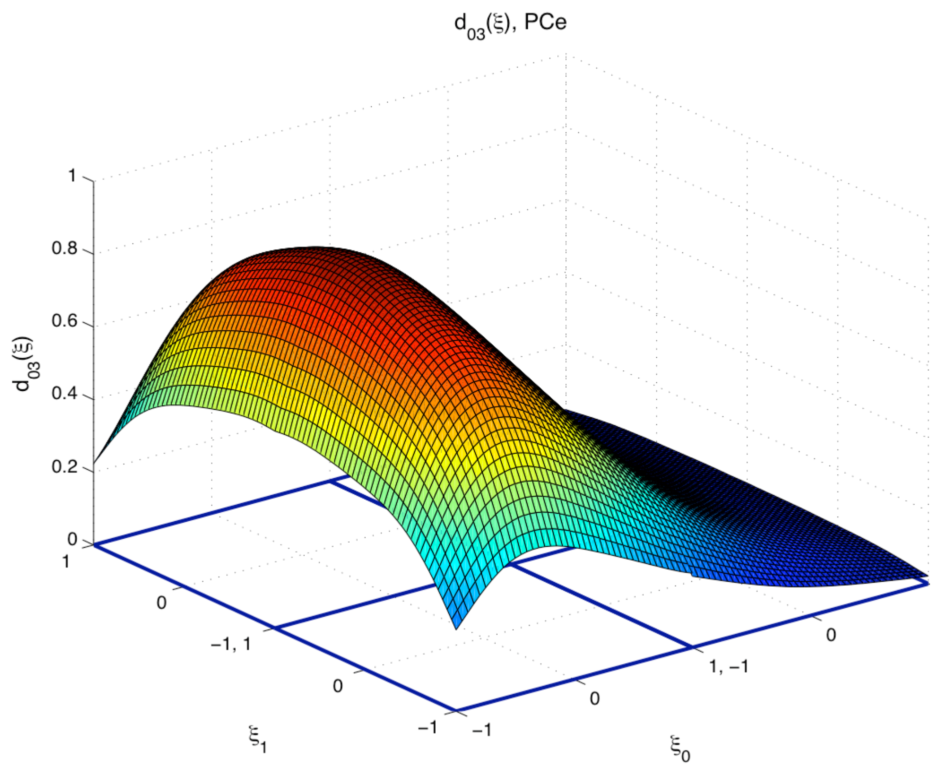
- Predicted value of the scalar field at *(x,y) = (0,0)*; ***t = 0.15***
- Convergence with respect to **order** *p*



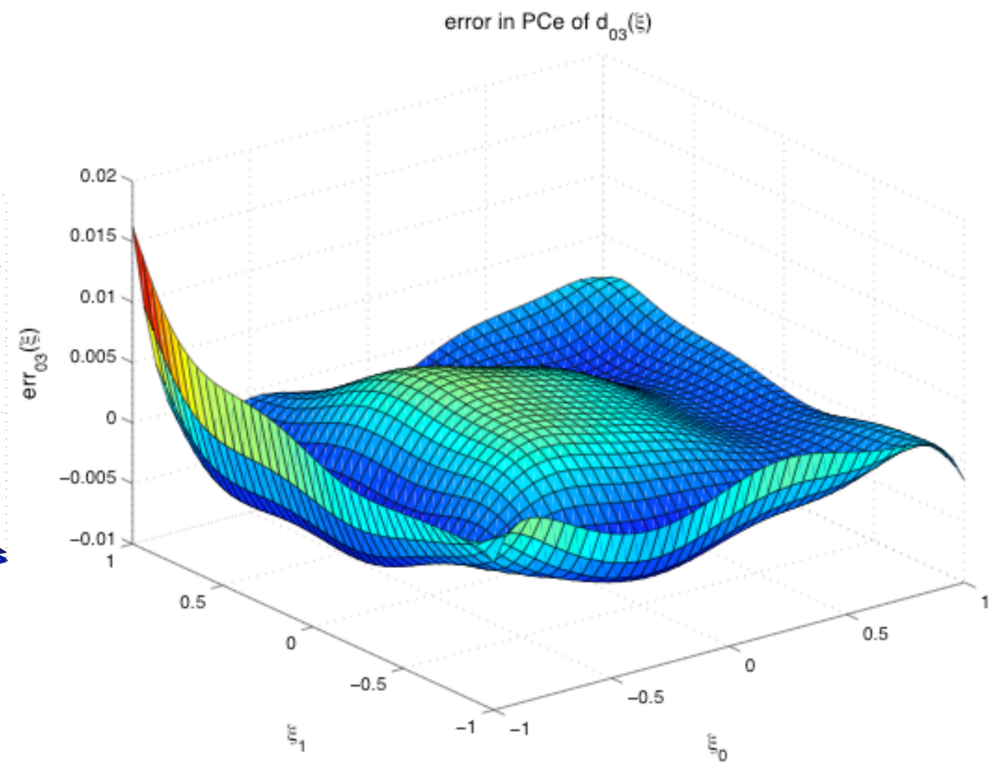
- Compare times: ***sensitivity*** information contained in the PCe…

Surface response and error

- Predicted value of the scalar field at *(x,y) = (0.0,0.5); **t = 0.15:***



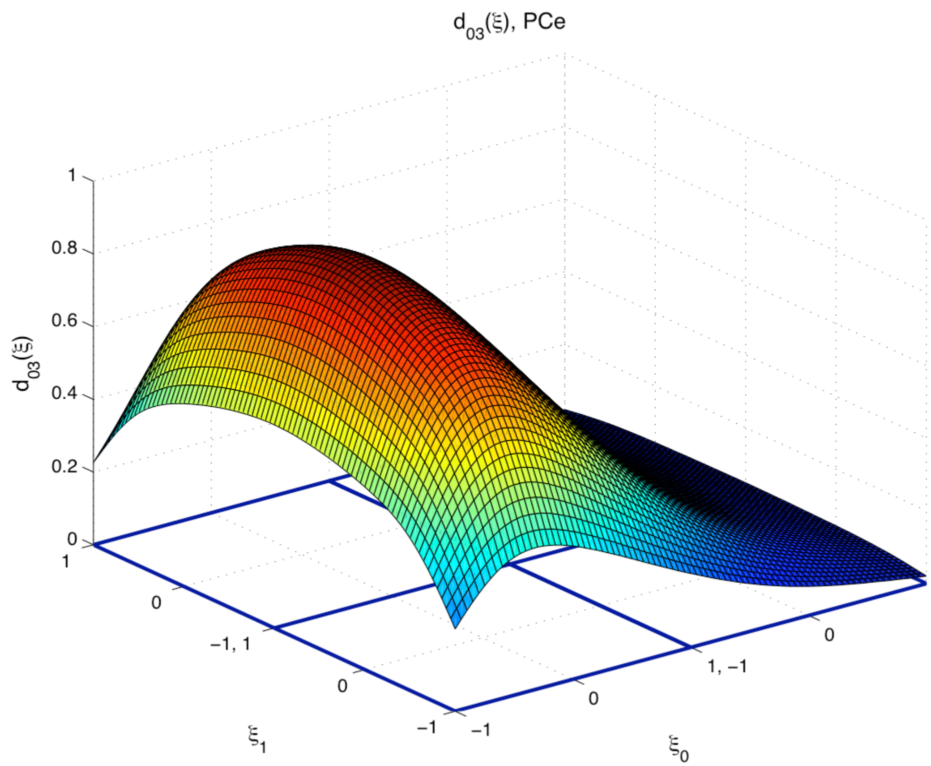
response $d_3(\xi)$ via PC (p=6)

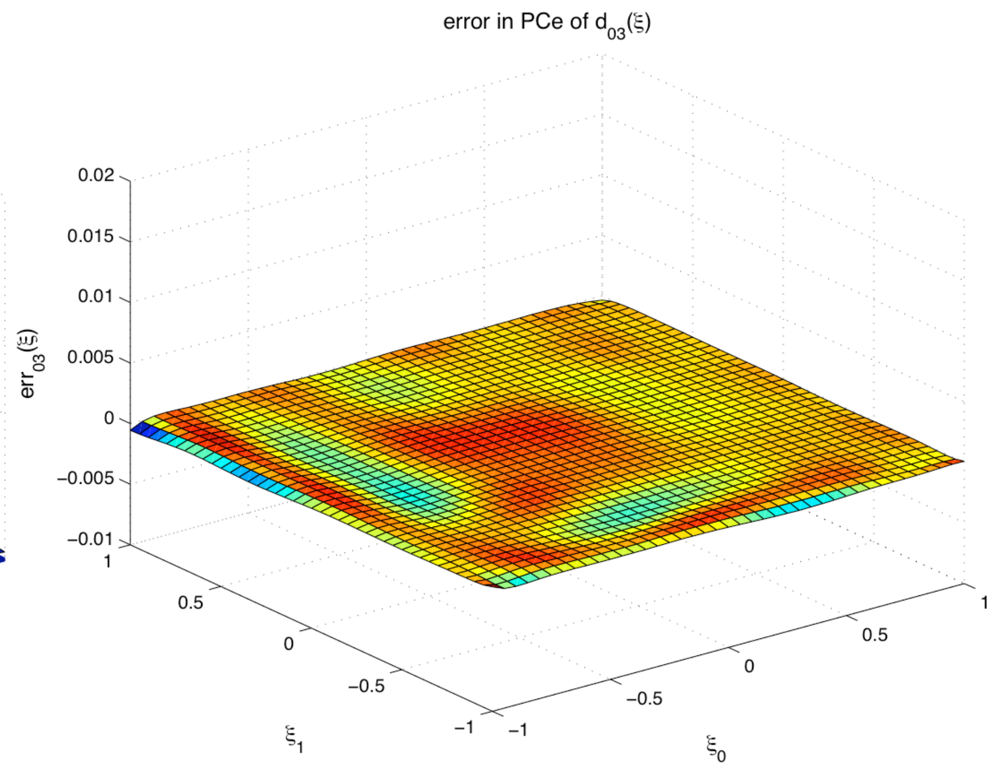


error: $d_3(\xi)$ - $G_3(m(\xi))$

Surface response and error

- Predicted value of the scalar field at *(x,y) = (0.0,0.5); **t = 0.15:***



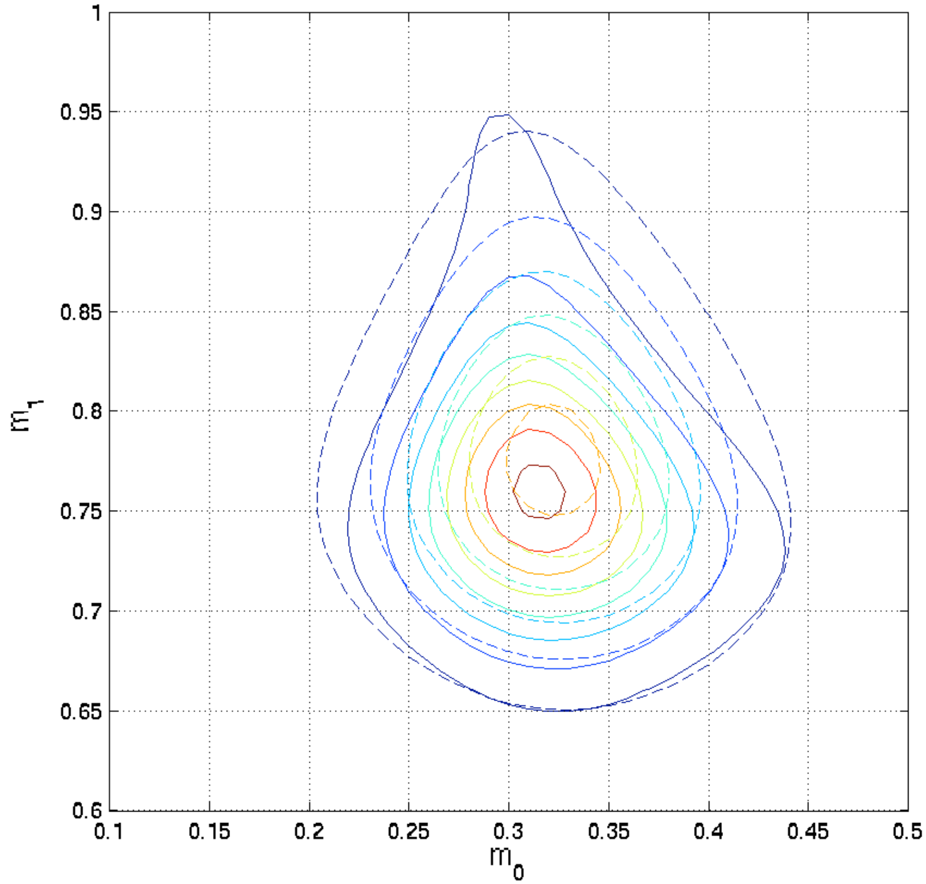


response $d_3(\xi)$ via PC (p=9)

error: $d_3(\xi)$ - $G_3(m(\xi))$

Posterior density

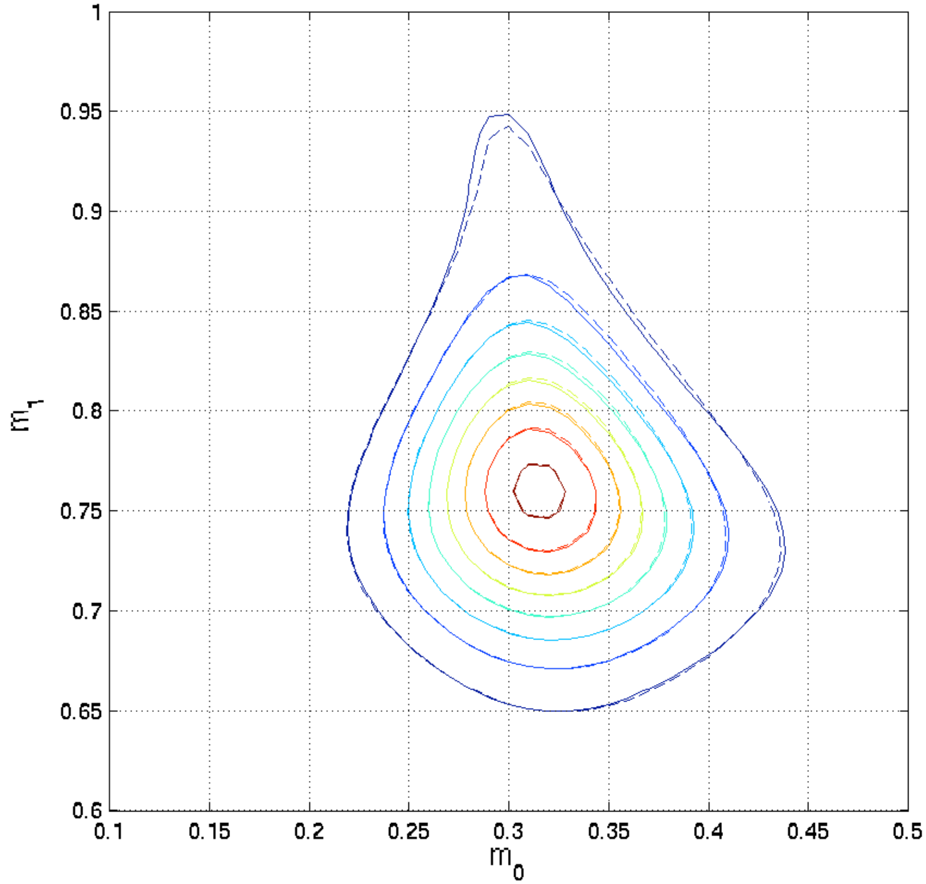
- 3×3 grid of sensors; measure at $t = \{0.05, 0.15\}$; $\boldsymbol{d}$ from noisy observations of a source at (x,y) = (0.25,0.75).



p=3 (dashed) vs direct (solid)

Posterior density

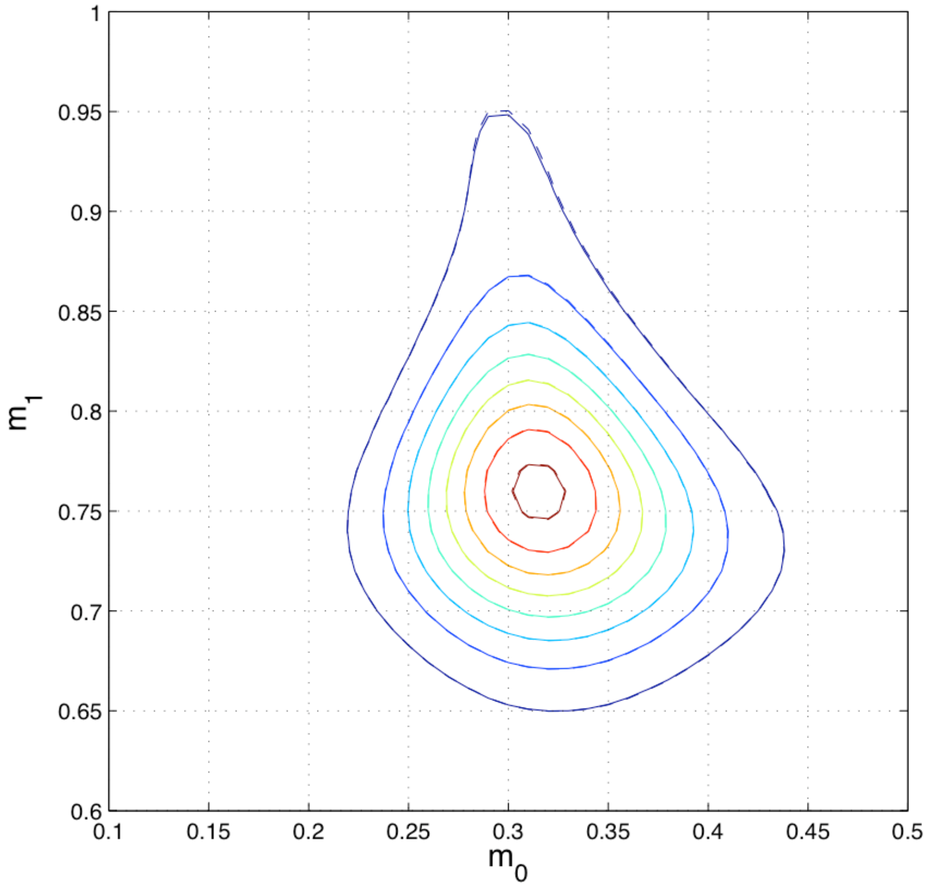
- 3×3 grid of sensors; measure at $t = \{0.05, 0.15\}$; $\boldsymbol{d}$ from noisy observations of a source at (x,y) = (0.25,0.75).



p=6 (dashed) vs direct (solid)

Posterior density

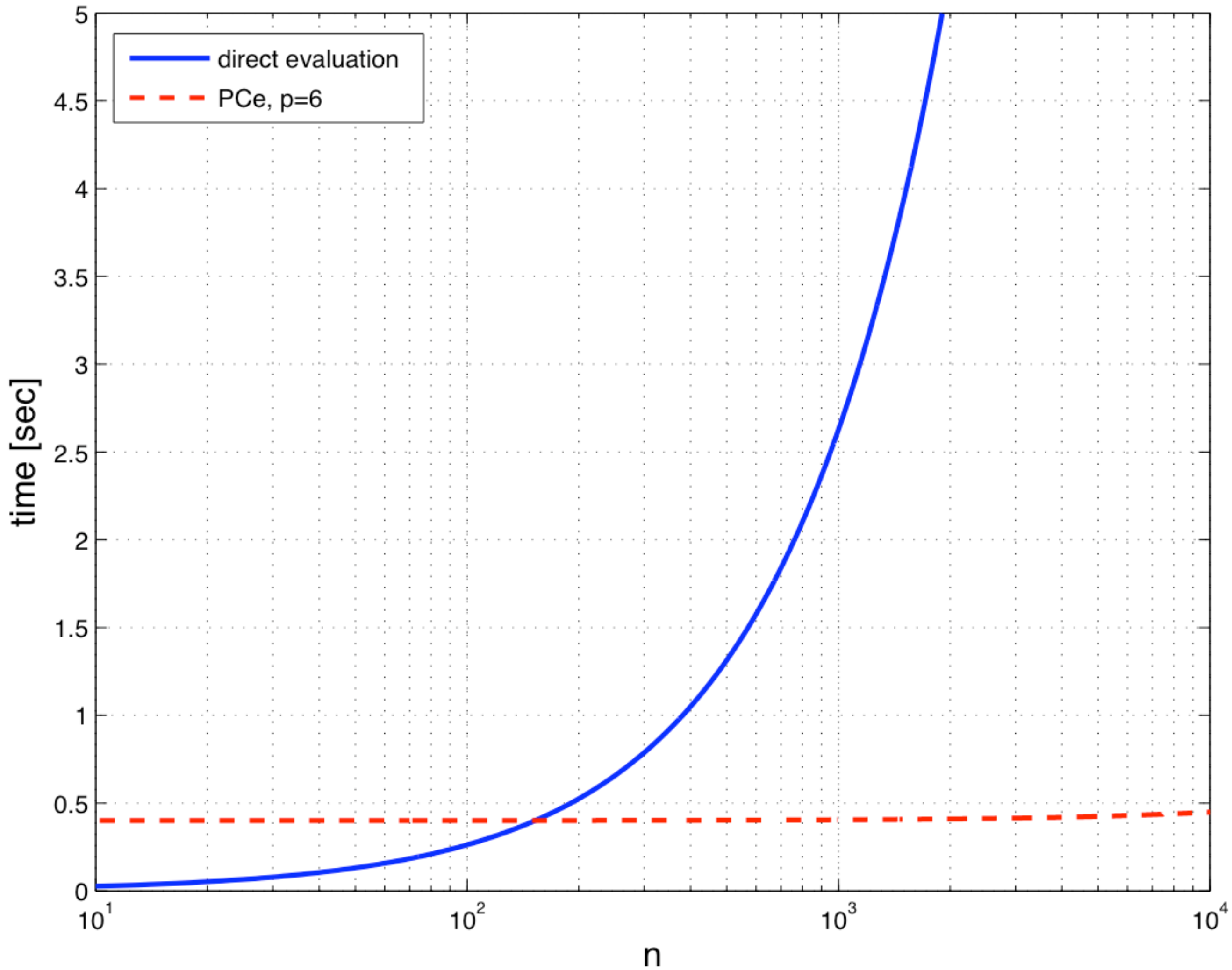
- 3×3 grid of sensors; measure at $t = \{0.05, 0.15\}$; ***d*** from noisy observations of a source at (x,y) = (0.25,0.75).



p=9 (dashed) vs direct (solid)

Monte Carlo speedup

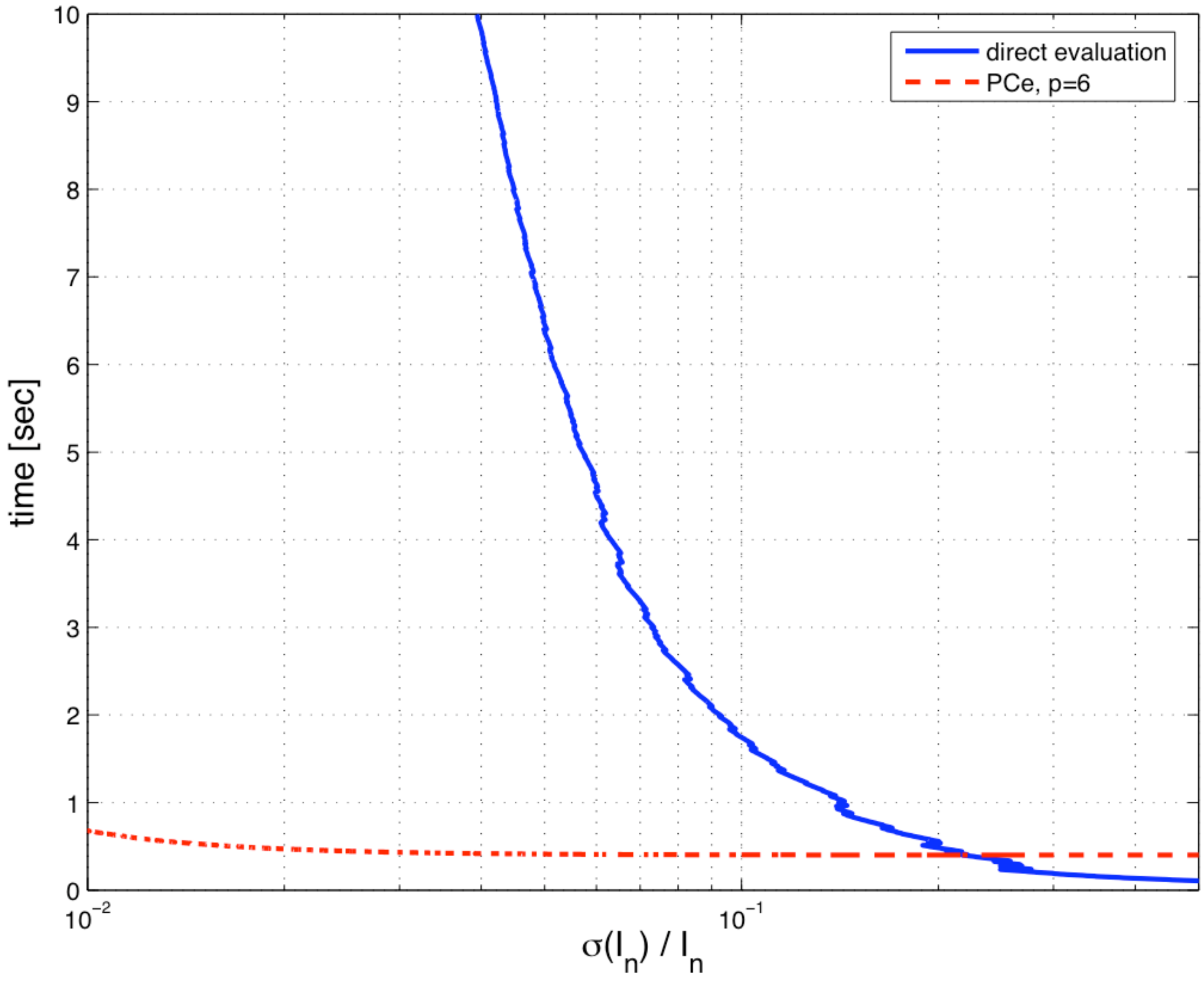
- Posterior mean: total computational time vs number of samples



Per-sample cost reduced by 2–3 orders of magnitude!!

Monte Carlo speedup

- TOTAL computational time vs relative standard error



$$Var[I_n] \rightarrow \frac{\sigma^2}{n} \quad \text{where}$$

$$\sigma^2 = Var_{p_m}[f(m)L(m)]$$

MCMC

- Construct a **Markov chain** of samples $m^{(t)}$ such that, after some burn-in period b, samples are being drawn from the posterior distribution $\pi(m)$
  - Markov chain defined by transition kernel $K\left(m^{(t+1)} \middle| m^{(t)}\right)$
  - π is the ***stationary distribution:*** $\int \pi(m) K\left(m^{(t+1)} \middle| m\right) dm = \pi(m)$
- How? [Metropolis 1953, Hastings 1970, Tierney 1995]
  - Proposal distribution $q\left(y \middle| m_t\right)$
  - Acceptance probability $0 < \alpha \leq 1$:

$$\alpha\left(m_t, y\right) = \min\left(1, \frac{\pi\left(y\right) q\left(m_t \middle| y\right)}{\pi\left(m_t\right) q\left(y \middle| m_t\right)}\right)$$

  - Acceptance $\Rightarrow$ $m^{(t+1)} = y$; otherwise $m^{(t+1)} = m^{(t)}$
- Ergodic average:

$$E\left[f\right] \approx \bar{f}_n \equiv \frac{1}{n-b} \sum_{t=b+1}^{n} f(m_t)$$

MCMC

- **Why use MCMC?**
  - Directly "simulate" the posterior— more efficient sampling
  - No normalization
  - Automatic marginalization

- Under certain conditions (*irreducibility*, *recurrence*)
  - SLLN: $\bar{f}_n \xrightarrow{a.s.} E_\pi[f]$
  - CLT: $\sqrt{n}\left(\bar{f}_n - E_\pi[f]\right) \xrightarrow{i.d.} N\left(0, \varsigma^2\right)$

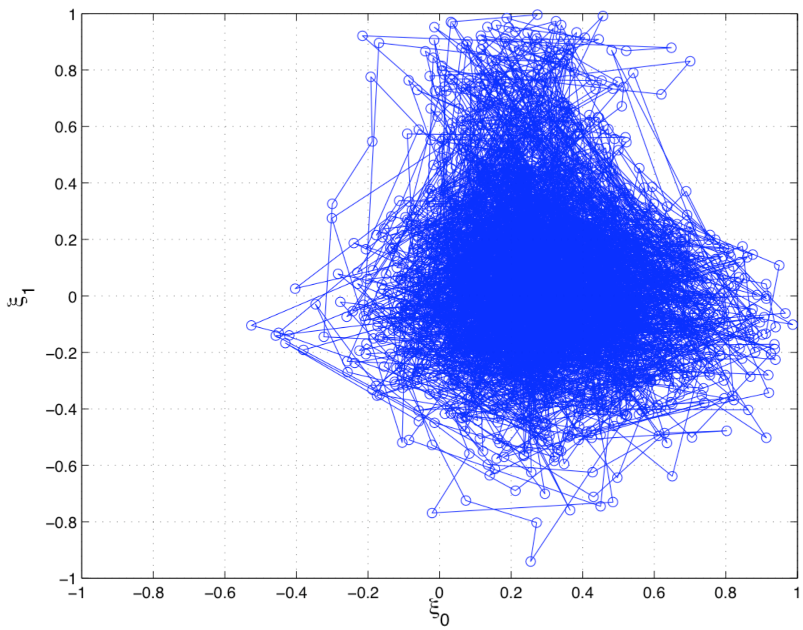
    $\varsigma^2 = \sigma^2 + 2\sum_{s=1}^{\infty}\gamma(s)$ where $\gamma(s) = E_\pi\left[\left(m^{(t)} - \langle m\rangle\right)\left(m^{(t+s)} - \langle m\rangle\right)\right]$

- For difficult distributions, diagnosing/verifying convergence still requires practical experience…

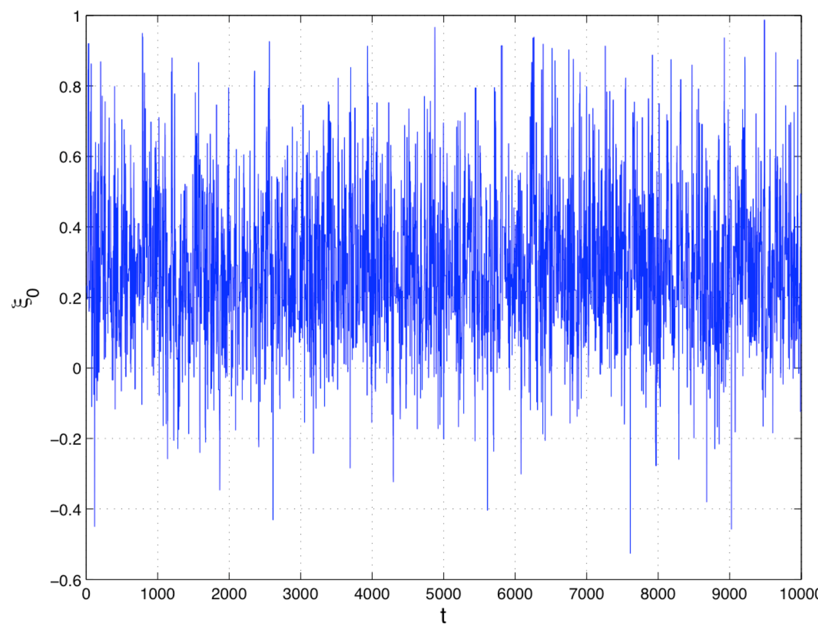
MCMC

- Apply random-walk Metropolis to the PC-transformed problem:

$$E_\pi[f] = \int_{\tilde{\Xi}} \underbrace{f\left(g(\xi)\right)}_{\tilde{f}(\xi)} \underbrace{\frac{L\left(g(\xi)\right) p_m\left(g(\xi)\right) \left|\det(Dg)\right|}{k}}_{\tilde{\pi}(\xi)} d\xi$$



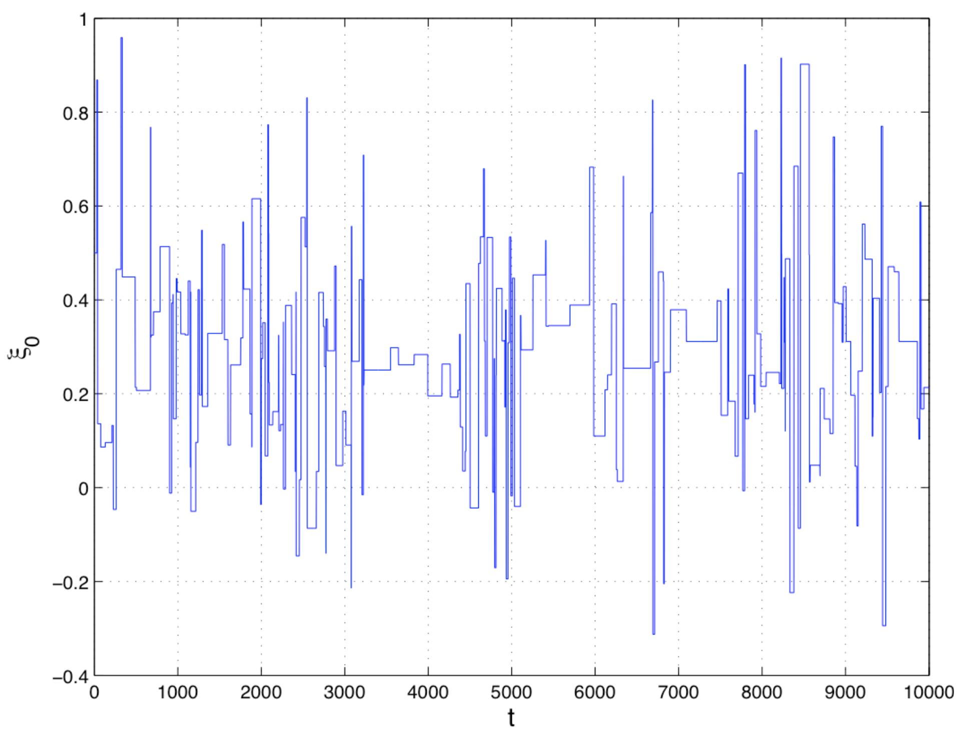
chain position in the ξ plane



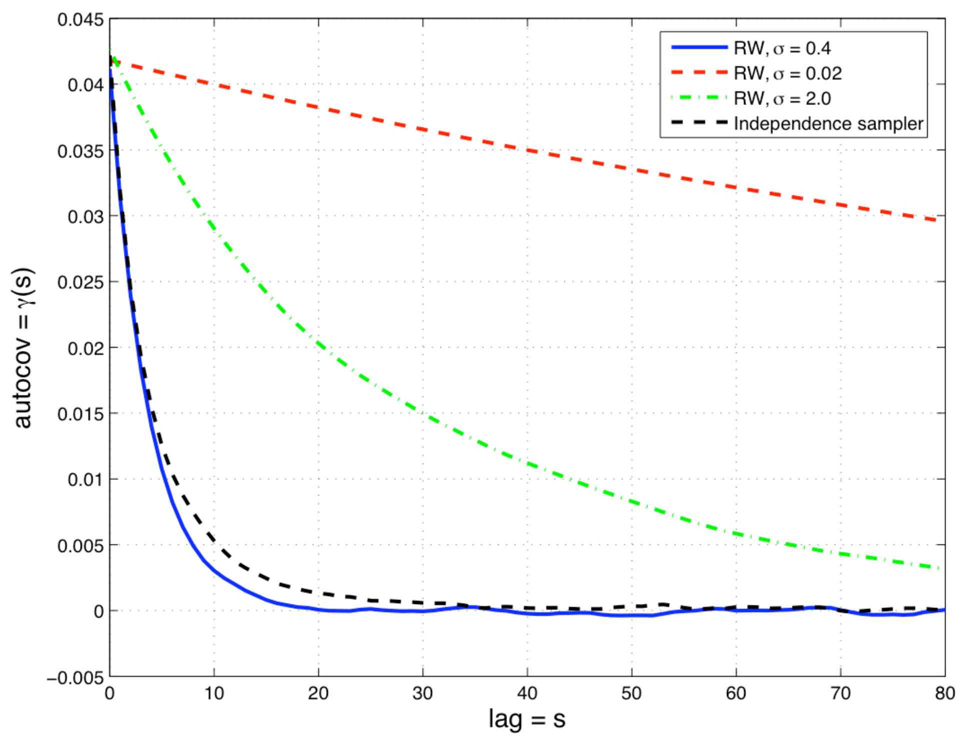
ξ_0-coordinate of chain position versus time

MCMC

- Mixing, good and bad:



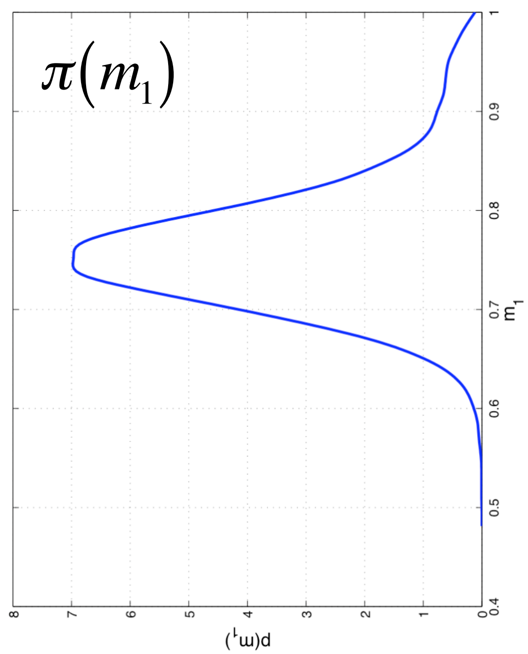
ξ_0–coordinate of the chain,
RWM with σ=2.0

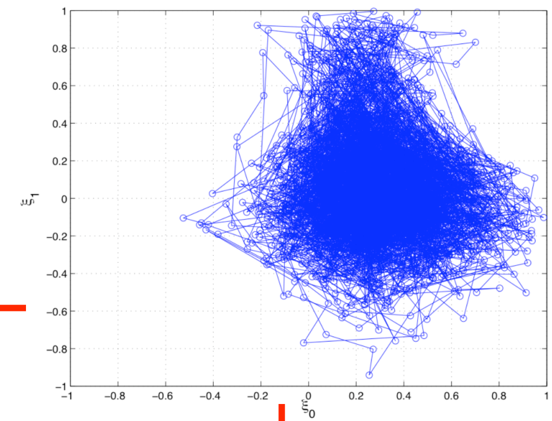


autocovariance for different samplers

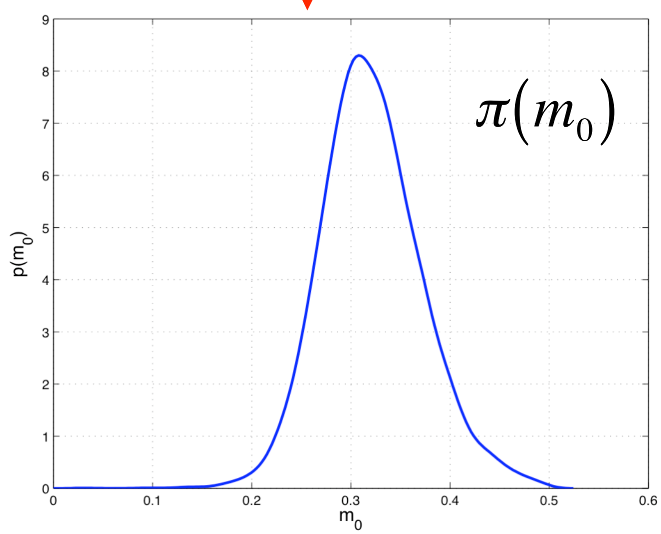
MCMC

- Marginal distributions via kernel density estimation:





samples from $\tilde{\pi}(\xi_0, \xi_1)$



Extending the Bayesian framework

- Inversion from forward models with *additional* parametric uncertainty (m_P):

$$d \approx G(m_I, m_P)$$

e.g., uncertain diffusivity

$$p(m_I, m_P | d) \propto p(d | m_I, m_P)\, p(m_I)\, p(m_P)$$

$$\int dm_P \longrightarrow \pi(m_I)$$

posterior on m_I

  - Propagate both $p(m_I)$ and $p(m_P)$ with PCe

- Uncertain forward models—another approach:

$$\boxed{F(\mathrm{d}|m_I)} \quad \rightarrow \quad L(m) = \int p_d(\mathrm{d}) F(\mathrm{d}|m_I)\, d\mathrm{d}$$

  - The exact forward model is now a special case:

$$F(d|m_I) = \delta(d - G(m))$$

Conclusions

- Bayesian inference for inverse problems
  - A complete approach to noisy data, incomplete data, ill-conditioning, and stochastic forward problems.
  - A quantitative description of *uncertainty* in the inverse result.
- Accelerating Bayesian inverse problem solutions with PCe:
  - Spectral representation of random variables; Galerkin projection.
  - Propagate prior uncertainty through forward model; *rapid* sampling by evaluating PCe
  - Choice of basis, order, decomposition of the prior support.
  - Sampling strategies (MC, MCMC)
- Demonstrate w/source inversion in transient diffusion

Ongoing work

- Larger problems, more complex source inversion:
  - **Multiple sources**, additional uncertain source parameters
  - PCe approaches for inverse problems on **continuous fields**
  - Add **convective transport!**

- Inverse problems in disease propagation
  (with J. Ray, K. Devine, P. Fast)

- Structural inference: building models of biological kinetic networks (e.g., gene regulatory networks from microarray data)